IJICT International Journal of Information & Communication Technology

ITRC

Volume 2- Number 2- July 2010 (pp.29-41)

A New GA Approach Based on Pareto Ranking Strategy for Privacy Preserving of Association Rule Mining

Mohammad Naderi Dehkordi
Faculty member of Islamic Azad University-Najafabad Branch
Isfahan, Iran
naderi@iaun.ac.ir

Kambiz Badie
Education and Research Institute for ICT
Tehran, Iran
k_badie@itrc.ac.ir

Ahmad Khadem-Zadeh
Education and Research Institute for ICT
Tehran, Iran
zadeh@itrc.ac.ir

Received: August 19, 2009- Accepted: July 25, 2010

Abstract—**With fast progress of the networks, data mining and information sharing techniques, the security of the privacy of sensitive information in a database becomes a vital issue to be resolved. The mission of association rule mining is discovering hidden relationships between items in database and revealing frequent itemsets and strong association rules. Some rules or frequent itemsets called sensitive which contains some critical information that is vital or private for its owner. In this paper, we investigate the problem of hiding sensitive knowledge. We decide to hide sensitive knowledge both in frequent patterns and association rule extraction steps. In order to conceal association rules and save the utility of transactions in dataset, we select Genetic Algorithm to find optimum state of modification. In our approach various hiding styles are applied in different multi-objective fitness functions. First objective of these functions is hiding sensitive rules and the second one is keeping the accuracy of transactions in dataset. After sanitization process we test the sanitization performance by evaluation of various criterions. Indeed our novel framework consists of dataset preprocessing, Genetic Algorithm-based core approach and different sanitizing measurements. Finally we establish some experiments and test our approach by larger datasets and compare the performance with well-known existing ones.**

Keywords- ***Association Rule Mining, Privacy Preserving, Sensitive Association Rule, Multi-objective Optimization, Genetic Algorithms***

I. Introduction

Rapid progress of storage and retrieval technology and key advances in size of storage media leads to a huge size of databases and couple of transactions. So the major need is how to use this size of information and extract their useful knowledge. Therefore new research area has been established as data mining to extract new hidden information from great size of data. Data mining has been adopted as a major part of knowledge discovery process. It has some built-in techniques to extract new information or verify hypothesis. These techniques include approaches for recognizing relationships that could take the structure of association rules, sequences, classifiers, etc. among others. The frequent itemsets mining process is an important fundamental step in large number of these activities, especially in association rules mining. Moreover, information sharing across organizations is one of the most important needs in inter-communication business. Transactional databases have also been widely recognized in the greater part of businesses. This information sharing increases the risk of sensitive relationship disclosure. On the other hand, latest progresses in data mining technology have increased the opportunity of such disclosure. As an in point case, think about a pharmaceutical company that asks its clients to disclose the diseases they have, in order to study the relationships in their occurrences. For instance, "Adult females with malarial infections are also prone to contract tuberculosis". While the company may be obtaining the data exclusively for legal data mining purposes that would ultimately reflect it-self in better service to the client, at the same time the client might worry that if her medical records are either inadvertently or deliberately exposed, it may adversely affect her employment opportunities [1].

Similar motivating instances are discussed in [2], [3], [4] and [5]. Privacy preserving in association rule mining methods try to provide a solution to this critical problem. These methods do so by accepting a small number of modifications in the original database that will forbid the creation of sensitive itemsets at a pre-specified support or confidence threshold usually set by the owner of the data. By preventing production of sensitive itemsets, we can be sure that at the given minimum support or confidence threshold, no sensitive association rule would be exposed. The modification process can affect the original set of rules that can be mined from the original database, either by hiding rules which are not sensitive (*lost rules*), or introducing rules in the mining of the modified database, which were not supported by the original database (*ghost rules*). We have tried to minimize these unpleasant results by performing minimum and appropriate modifications in original dataset.

In this paper we propose a novel framework based on genetic algorithms for privacy preserving of association rule to find the best solution for sanitizing original dataset based on multi-objective optimization. We involve balancing some critical factors in database sanitization; Starting from some changes to hide sensitive association rules and do not so many changes to loss non-sensitive association rules and finally some changes in such a way that no spurious association rules would be extracted. We try to satisfy all of these objectives simultaneously. There are so many methods to solve multi-objective problems. Some of the most well-known methods are: weighted sum strategy, ε-constraint method, set of non-inferior solutions (Pareto frontier) and goal attainment method. In our framework we try to solve this optimization problem by Pareto ranking strategy in genetic algorithm.

The rest of paper is organized as follows: Section 2 gives a summary of the high-tech methodologies and related works for privacy preserving in data mining and association rule hiding with dataset sanitization. In Section 3 we describe problem formulation and enlighten the major concepts upon which we base the proposal for the new privacy preserving framework. Section 4 describes our proposed solution for dataset sanitization against association rule mining. Section 5 presents the experiments we performed first in case study and second in large scale datasets to introduce our approach and to prove the effectiveness of our method. Finally the conclusion will be given in Section 6.

II. Related Works

Privacy issue of data management has been focused for long time. For example one of earlier papers in this research area was by Atallah et al. [6]. In this work proved that many of underlying problem in privacy preserving are NP-Hard. Therefore, most of researches have done in heuristic approach. Some of these works are stated as follows. In one of the latest papers by Verykios et al. [7], has addressed the problem of privacy preserving in association rules as "hiding association rules" and they have done by heuristic approaches. Because of many underlying NP-hard problems [6], using heuristic approaches is not astonishing. Some of most important works have done on hiding of frequent itemsets [8]. Although they proposed four approaches to preserve privacy in datasets, these approaches are relatively limited as like as other related heuristic based works and do not warranty global optimality of their solutions in sanitization problem (this is a major drawback of heuristic approaches). Wang et al. [14] propose a heuristic approach that achieves to fully eliminate all the sensitive inferences, while effectively handling overlapping rules. Their proposed algorithm identifies the set of attributes that influence the existence of each sensitive rule the most and removes them from those supporting transactions that affect the non-sensitive rules the least. Wang and Jafari [15] propose two modification schemes that incorporate "unknowns" and aim at the hiding of predictive association rules, i.e. rules containing the sensitive items on their LHS. Both algorithms rely on the distortion of a portion of the database transactions to lower the confidence of the association rules. Amiri [13] proposes three effective, multiple rule hiding heuristics that surpass SWA by offering higher data utility and lower distortion, at the rate of computational cost. Although there is similarity between these approaches, the proposed schemes do a better job in modeling the overall objective of a rule hiding algorithm. The work of Abul et al. [18] is the first to concentrate on the NP-hardness issue involving the optimal hiding of sequences and to provide a heuristic, polynomial time algorithm that carries out the sanitization task. A different research direction concerns the use of database reconstruction approaches. Prominent research efforts towards this direction include the work of several researchers in the field of inverse frequent itemset mining [16, 17, 21]. Saygin [20] extends the protocol-based approaches to capture the clustering of spatio-temporal data. The proposed protocol is in compliance with a series of trajectory comparison functions and allows for secure similarity computations through the use of a trusted third party. Gkoulalas and Verykios [19] propose an exact approach for hiding sensitive rules that uses the itemsets belonging in the revised positive and the revised negative borders to identify the candidate itemsets for sanitization.

In this article we have tried to find optimal solutions for sanitization problem. One of the most important issues in sanitization problem is that there are different criterions in privacy preserving and it can not realistic that all of these criterions are satisfied at the same time. On the other hand, in sanitization of dataset tried to keep all of these measurements at best level. In this paper we have tried to solve this multi-objective optimization problem by appropriate Genetic Algorithm approach with proper Pareto frontier fitness function. Indeed we have supposed that there are no specified priorities or costs as weights for the objectives and finally showed a set of *non-dominated* solutions (Pareto frontier) as result.

III. Problem Formulation

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items and let D is the dataset of transactions that contains sensitive information and it should be sanitized before publishing. Itemset denoted as $X \subseteq I$. Each itemset which contains *k* items called *k-itemset.* Let $D = \{T_1, T_2, ..., T_n\}$ be a set of transactions. The well known measure in frequent itemset mining is *support* of itemset. The *support* measure of an item $X \subseteq I$ in database D, is the count of transactions contain X and denoted as *Support_count(X)*. An itemset *X* has *support* measure *s* in dataset D if *s%* of transactions support *X* in dataset *D*. *Support* measure of *X* is denoted as *Support(X)*.

$$Support(X) = \frac{Support\_count(X)}{n} \times 100$$

where *n* is number of transactions in dataset *D*.
Itemset *X* is called frequent itemset when Support(X)$\geq MST$, where *MST* is an acronym for "Minimum Support Threshold" that is predefined threshold. After mining frequent itemsets, the association rule is an implication of the form $X \rightarrow Y$, where $X, Y \subset I$ and $X \bigcap Y = \phi$. The *Confidence* measure for rule $X \rightarrow Y$ in dataset *D* is evaluated as follows:

$$Confidence(X \rightarrow Y) = \frac{Support(XY)}{Support(X)} \times 100.$$

Note while the *support* is a measure of the frequency of a rule, the *confidence* is a measure of the strength of the relation between sets of items. Association rule mining algorithms scan the dataset of transactions and evaluate the *support* and *confidence* of candidate rules to determine if they are considerable or not. A rule is considerable if its *support* and *confidence* is higher than the user specified minimum support and minimum confidence threshold. In this way, algorithms do not retrieve all possible association rules that can be derivable from a dataset, but only a very small subset that satisfies the minimum support and minimum confidence requirements set by the users. An association rule-mining algorithm works as follows. It finds all the sets of items that appear frequently enough to be considered relevant and then it derives from them the association rules that are strong enough to be considered interesting. The major goal here is to prevent some of these rules that we refer to as "sensitive rules", from being revealed. The problem of privacy preserving in association rule mining (so called association rule hiding) focused on this paper can be formulated as follows:

Given a transaction database D, minimum support threshold “MST”, minimum confidence threshold “MCT”, a set of significant association rules R mined from D and a set of sensitive rules $R_{Sen} \subseteq R$ to be hided, generate a new database D', such that the rules in $R_{non-Sen} = R - R_{Sen}$ can be mind from D' under the same “MST” and “MCT”. Further, no normal rules in $R_{non-Sen}$ are falsely hidden (lost rules), and no extra spurious rules (ghost rules) are mistakenly will mined after the rule hiding process.

In [6] proved that solving above problem by sinking the support of the large itemsets via removing items from transactions or adding fake item into the transactions (also referred to as “sanitization” problem) are an NP-hard problem. Therefore, we are looking for a special modification of *D* (the source dataset) in D' (sanitized dataset which is going to be released) that *maximizes* the number of rules in $R_{non-Sen}$ (*minimizing* number of lost rules) that can still be mined. Therefore we involve specific optimization problem. In one side we must conceal the sensitive association rule, thus it is necessary to modify the dataset and in the other side we should keep the utility of modified dataset to extract useful information and rules. In order to solve this optimization problem we have developed a framework and some criterion to evaluate our sanitization performance.

IV. The Proposed Solution

In the following section we will explain our approach specifically. The critical phase in this work is "preprocessing phase" and the related specifications of the fitness function which is to be used for our Genetic Algorithm method.

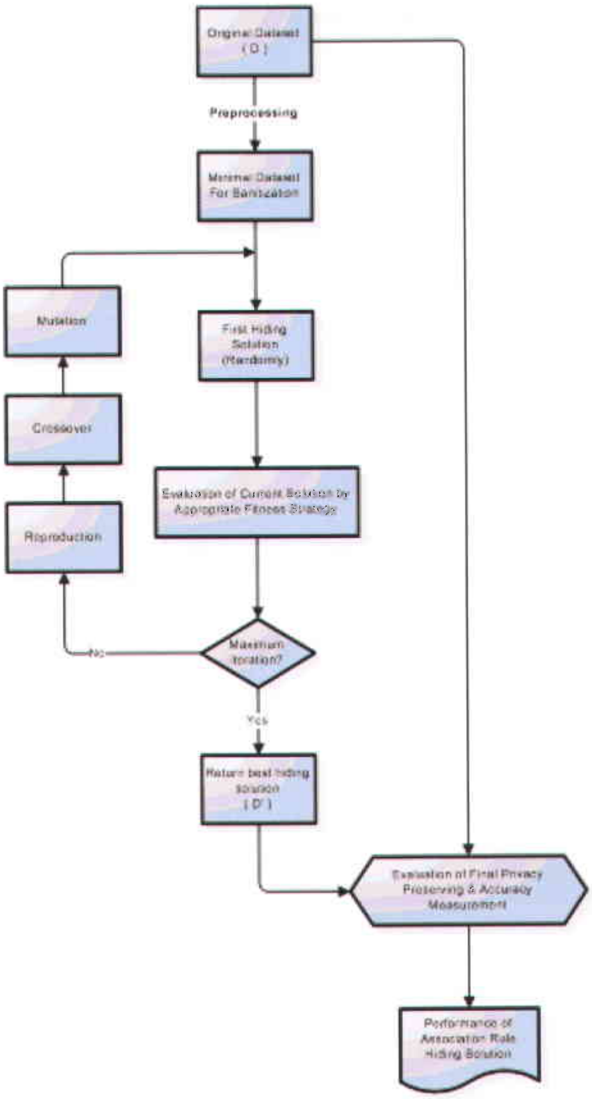

Figure 1. Main phases in our proposed framework

A. *Preprocess of Original*

B. *Dataset*

The overall workflow in our approach is depicted in Figure 1. The whole approach is divided into two phases: 1-Preprocessing of original dataset 2-Searching for the best sanitization solution based on Genetic Algorithm and according to appropriate fitness strategy in minimal dataset.

The first phase is to preprocess of original dataset and address minimal itemsets that need modification. We propose two strategies in preprocessing of the dataset. First, we can select all transactions that support sensitive itemsets. In this strategy a common item(s) between the transaction and sensitive rule is required to select the transaction. Therefore in this strategy each transaction that has sensitive items is addressed to change. So we should have amount of locations that possibly changed either fully support or partially support sensitive association rule. As a result, we need more space to generate longer chromosomes and manipulation of these chromosomes needs more time. Further, we may have so many candidate locations for modifications in original dataset, and the utility of dataset may be affected more.

```
INPUT: a set of sensitive association rules to hide $R_{sen}$ and
            original dataset $D$

OUTPUT: the minima dataset for modification $D_{Minimal}$
Begin
$D_{Minimal} \leftarrow \phi$    // minimal dataset is empty in the beginning
for each sensitive association rule $r \in R_{sen}$ do {
  for each transaction $t \in D$ do {
    $common\_items \leftarrow (r_{items} \cap t_{items})$
```

```
        if $common\_items \neq \phi$ then            //Phase1:
    Select transaction t that  partially supports sensitive rule r
            append_to_dataset ( $D_{Minimal}, t_{common\_items}$ );
        //Phase2: Select sensitive items form transaction t
      }
  }
  End
```

Figure 2. Preprocessing algorithm based on partially supported transactions strategy.

```
INPUT: a set of sensitive association rules to hide $R_{sen}$ and
              original dataset $D$

OUTPUT: the minima dataset for modification $D_{Minimal}$
Begin
$D_{Minimal} \leftarrow \phi$
// minimal dataset is empty in the beginning
for each sensitive association rule $r \in R_{sen}$ do {
  for each transaction $t \in D$ do {
    $common\_items \leftarrow (r_{items} \cap t_{items})$
    if $common\_items = r_{items}$ then
          //Phase1: Select transaction t that  fully supports
sensitive rule r
          append_to_dataset ( $D_{Minimal}, t_{common\_items}$ );
          //Phase2: Select sensitive items form transaction t
  }
}
End
```

Figure 3. Preprocessing algorithm based on selection of fully supported transactions strategy.

On the other hand, in this strategy we make more changes and the sensitive items will be concealed by more scrupulosity. The algorithm of first preprocessing strategy is depicted in Figure 2. Second, we can use minimum confidence threshold to select all transactions that support sensitive association rules. In this strategy each transaction that fully supports the sensitive association rule are addressed to change. In comparison with the first strategy, the strategy candidates a fewer number of transaction to change. Because many of the transactions do not support the whole typical sensitive association rule. On the other hand, in this strategy we modify a smaller number of transactions. Hence, accuracy and usefulness of dataset is also maintained. The algorithm of the second preprocessing strategy is depicted in Figure 3.

Therefore, if the high priority goal is to fully preserve sensitive items, we should select first preprocess strategy and if we are going to maintain utility of dataset more than before, the second strategy is a better choice for preprocessing. The overall view of preprocessing phase is depicted in Figure 4.

C. GA Proposed Solution for Privacy Preserving

1) Genetic Algorithm Background

A Genetic Algorithm performs fitness tests on new structures to select the best population. Fitness determines the quality of the individual on the basis of the defined cost function. Genetic Algorithms are meta-heuristic search methods that have been developed by John Holland in 1975. [9,10] GA's applied natural selection and natural genetics in artificial intelligence to find the globally optimal solution to the optimization problem from the feasible solutions. In nature, an individual's fitness is its ability to pass on its genetic material. The fortune of an individual chromosome depends on the fitness value; the better the fitness value, the better the chance of survival. Genetic Algorithms solve design problems similar to that of natural solutions for biological design problems [11].

2) Population Generation and Chromosome Presentation

In Genetic Algorithms, a population consists of a group of individuals called chromosomes that represent a complete solution to a defined problem. Each chromosome is a sequence of 0s or 1s. The initial set of the population is a randomly generated set of individuals. A new population is generated by two methods: steady-state Genetic Algorithm and generational Genetic Algorithm. The steady-state Genetic Algorithm replaces one or two members of the population; whereas the generational Genetic Algorithm replaces all of them at each generation of evolution. In this work a generational Genetic Algorithm is adopted as population replacement method. In this method tried to keep a certain number of the best individuals from each generation and copies them to the new generation (this approach known as elitism).

Each transaction is represented as a chromosome and presence of an ith item in transaction showed by 1 and absence of the item by 0 in ith bit of transaction. The fitness of a chromosome is determined by several factors and different strategies. Each population consists of several chromosomes and the best chromosome is used to generate the next population. For the initial population, a large number of random transactions are chosen. Based on the survival fitness, the population will transform into the future generation.

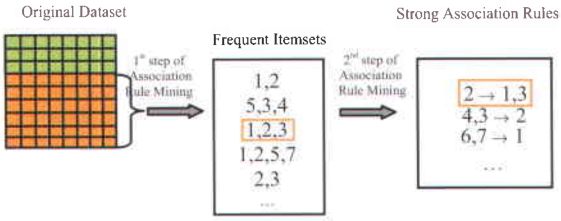

Figure 4. Association Rule Mining Phases

3) *Fitness Strategies*

The dynamic area in this research is multi-objective optimization. The idea is quite simple. In these strategies fitness measurements happen in two stages. In stage one, each objective is measured with its natural fitness measurement (as like as weighted sum approach). However, these scores are not merged at all, but are kept separate for each population member within a *vector* of scores. A Genetic Algorithm would therefore evaluate each individual according to all the multi-objective evaluation tests as are necessary for the problem. Stage two involves finding overall rankings for the population. Recall that ranked fitness measurements discard absolute fitness scores, and instead replace them with integer numbers (1, 2, 3,..., with 1 being the most fit, 2 being 2nd fittest, etc.). The ranking done here uses the Pareto ranking strategy. The idea behind Pareto ranking is that it will never try to compare quantity of two objectives in different types: each dimension of the problem is always kept independent of the other dimensions, and an individual is better than, or *dominates*, another individual if it is shown to be at least as good in all dimensions, and better in at least one dimension. For a minimization problem (one in which we are trying to minimize scores), then for two individuals *U(u(1), u(2), ..., u(k))* and *V (v(1), v(2), ..., v(k))*, we say that:

$$U \text{ dominates } V \quad \textit{iff:}$$
$$\forall i : u(i) \le v(i) \wedge \exists i : u(i) < v(i)$$

The first expression with "for all" says that there is *U* is at least as good as *V* is in all aspects. And the second expression ("there exists") says that there is at least one aspect of *U* that is definitely better than *V*. Therefore it is so clear that *U* is superior to *V*, because it is better in at least one aspect, and not worse in any aspect.

The Pareto ranking algorithm relies on the idea of domination. It first goes through the entire population (all sanitization solutions for this problem) to find the non-dominated individuals (superior sanitization solutions). These are the individuals in which nothing dominates them. These will be assigned rank one (first one in ranking), the fittest individuals in the population. The ranking algorithms takes an individual A, and then looks through the rest of the population to see if any individual B dominates A. If so, then A cannot be in rank one, and it is skipped. If however, it is found that there is no B that dominates A, then A is assigned rank one. Once the entire population is evaluated for the rank one members, these rank one individuals are marked as "processed", and the whole procedure is repeated on the remaining population to find the rank two individuals... those that are *non-dominated* by any yet unranked individuals. This repeats until the entire population is assigned a rank.

The end result of the Pareto ranking is that each member of the population has a single Pareto rank value assigned to it. The lower rank, the better individual. These ranks can then be converted to a Roulette wheel or used within a tournament selection to create the next generation [12].

There will usually be sets of individuals in each rank as well. The individuals in a rank dominate all the individuals with higher rank numbers, and are in turn dominated by the sets with lower ranks. However, individuals in the same rank set are incomparable, in the sense that none of them is clearly better or worse than any other member of that set. Each individual will be better in some dimensions of the problem, but worse in others.

Based on Pareto ranking strategy, we have conducted four fitness evaluation strategies in this paper. We will discuss these strategies in following sections.

a) *Confidence-based Fitness Strategy*

First fitness strategy relies on both hiding all sensitive rules and minimum number of modification in original dataset. We design this fitness strategy based on Pareto ranking strategy as follows:

minimize *objective_1=Rules Hiding Distances*
AND
minimize *objective_2=Number of Modifications*

where:

- *Rules Hiding*

$$\textit{Distances} = \sum_{i=1}^{\textit{Number of sensitive Rules}} \textit{Rule}_i \;\; \textit{Hiding}$$

Distance

- $\textit{Rule}_i$ *Hiding Distance*

$$= \begin{cases} 0 & \textit{if Confidence}(\textit{Rule}_i) \le \textit{MCT} \\ \textit{Confidence}(\textit{Rule}_i) - \textit{MCT} & \textit{otherwise} \end{cases}$$

-
- *Number of*

$$\textit{Modifications} = \sum_{j=1}^{|\textit{Critical Transactions}| \times |I|} D'_j \oplus D_j$$

Where: $|\textit{Critical Transactions}|$ is number of critical transactions (in Figures 2 colored by orange) and $|I|$ is number of items in original database (denoted by *D*). And finally D'_j and D_j are j^{th} item of each dataset after and before sanitization respectively.

Association rule mining process depicted in Figure 2. In this fitness strategy we are trying to filter sensitive rules in 2nd step of mining process. Further, this strategy tried to apply minimum modifications in original dataset.

b) Support-based Fitness Strategy

Second fitness strategy relies on both hiding all sensitive itemsets and minimum number of modification in original dataset. We design this fitness strategy based on Pareto ranking strategy as follows:

minimize *objective_1= Itemsets Hiding Distances*
AND
minimize *objective_2= Number of Modifications*
where:

- $Itemset\ Hiding\ Distances = \sum_{i=1}^{Number\ of\ sensitive\ Itemsets} Itemset_i\ Hiding\ Distance$
- $Itemset_i\ Hiding\ Distance$

$$= \begin{cases} 0 & if\ Support\,(Itemset_i) \leq MST \\ Support\,(Itemset_i) - MST & otherwise \end{cases}$$

- $Number\ of\ Modifications = \sum_{j=1}^{|Critical\ Transactions| \times |I|} D'_j \oplus D_j$

Where: $|Critical\,Transactions|$ is number of critical transactions (in Figure 2 colored by orange) and $|I|$ is number of items in original database (denoted by *D*). And finally D'_j and D_j are jth item of each dataset after and before sanitization respectively.

In this fitness strategy we are trying to filter sensitive itemsets in 1st step of mining process (showed in Figure 2). Further, this strategy tried to apply minimum modifications in original dataset.

c) Hybrid Fitness Strategy

Third fitness strategy relies on hiding all sensitive rules and items. Further, minimum number of modification in original dataset is applied. We design this fitness strategy as hybrid of first and second strategies.

minimize *objective_1= Total Hiding Distances*
AND
minimize *objective_2= Number of Modifications*

where:

- $Total\ Hiding\ Distances = \sum_{i=1}^{Number\ of\ sensitive\ Itemsets / Rules} Itemset_i\ Hiding\ Distance + Rule_i\ Hiding\ Distance$
- $Itemset_i\ Hiding\ Distance$

$$= \begin{cases} 0 & if\ Support\,(Itemset_i) \leq MST \\ Support\,(Itemset_i) - MST & otherwise \end{cases}$$

- $Rule_i\ Hiding\ Distance$

$$= \begin{cases} 0 & if\ Confidence\,(Rule_i) \leq MCT \\ Confidence\,(Rule_i) - MCT & otherwise \end{cases}$$

- $Number\ of\ Modifications = \sum_{j=1}^{|Critical\ Transactions| \times |I|} D'_j \oplus D_j$

Where: $|Critical\,Transactions|$ is number of critical transactions and $|I|$ is number of items in original database (denoted by *D*). And finally D'_j and D_j are j^{th} item of each dataset after and before sanitization respectively.

In this fitness strategy we are trying to filter sensitive itemsets/rules both in 1st and 2nd steps of mining process (showed in Figure 2). Further, this strategy tried to apply minimum modifications in original dataset.

d) Min-Max Fitness Strategy

Fourth fitness strategy relies on minimizing number of sensitive rules and maximizing number of non-sensitive association rules that can be extracted from sanitized dataset. (See Figures 1 to 4 again). We design this fitness strategy as follows:

minimize *objective_1=* $|R' \cap R_{Sen}|$
AND
maximize *objective_2=* $|R' \cap R_{non-Sen}|$
or
minimize *objective_1=* $|R' \cap R_{Sen}|$
AND
minimize *objective_2=* $-|R' \cap R_{non-Sen}|$

where: $|R' \cap R_{Sen}|$ is number of sensitive association rules that is mined from sanitized dataset and $|R' \cap R_{non-Sen}|$ is number of non-sensitive association rules that is mined from sanitized dataset.

In this strategy tried to balance hiding all sensitive rules and keeping non-sensitive information. In other words, we have tried to preserve the privacy and accuracy of original dataset, simultaneously.

4) Selection

After evaluation of population's fitness, the next step is chromosome selection. Selection embodies the principle of "survival of the fittest". Satisfied fitness chromosomes are selected for reproduction. Poor chromosomes or lower fitness chromosomes may be selected a few or not at all. In this paper we have used Pareto ranking strategy. The end result of the *Pareto ranking* is that each member of the population has a single Pareto rank value assigned to it. The lower the rank, the better the individual. These ranks can then be converted to a "*Roulette-wheel*" or used within a "*Tournament*" selection to create the next generation. In *Tournament* selection, which is used in this paper, two chromosomes are chosen randomly from the population. First, for a predefined probability *p*, the more fit of these two is selected and with the probability *(1-p)* the other chromosome with less fitness is selected [19].

5) Crossover

Main function of crossover operation in Genetic Algorithms is combination two chromosomes together to generate new offspring (child). Crossover occurs only with some probability (crossover probability). Chromosomes are not subjected to crossover remain unmodified. The intuition behind crossover is exploration of new solutions and exploitation of old solutions. Better fitness chromosomes have a prospect to be selected more than the worse ones, so good solution always alive to the next generation. There are different crossover operators that have been developed for various purposes. Single-point crossover and multi-point are the most famous operators. In this paper single-point crossover has been applied to make new offspring. Normally high value of crossover probability is used (between 0.80 and 0.90).

6) Mutation

After performing crossover operation, the new introduced generation will only have the character of the parents. This behavior can lead to a problem where no new genetic material is introduced in the offspring and finding better population has been stopped. Mutation operator permits new genetic patterns to be introduced in the new chromosomes (random changed in random gene of chromosome). Mutation introduces a new sequence of genes into a chromosome but there is no guarantee that mutation will produce desirable features in the new chromosome. The selection process will keep it if the fitness of the mutated chromosome is better than the general population, otherwise, selection will ensure that the chromosome does not live to mate in future. Same as crossover operator, the mutation rate (mutation probability) is defined to manage how often mutation is applied. Contrasting crossover, the mutation rate is very low, about 0.005 to 0.01.

V. Performance Evaluation

To illustrate our proposed approach for the association rule hiding problem, validation of its feasibility and discussion about sanitization performance, let us consider an example.

A. Case Study

In this example we have original dataset and some sensitive association rule (See tables 1 to 3). Original dataset has shown in table 1 and the sensitive association rule in table 2. Before any modification in original dataset and with MST=0.33 and MCT=0.7, we can extract some association rules that are depicted in table 3. The specifications of our Genetic Algorithm for privacy preserving in association rule mining is showed in table 4.

TABLE I. Original dataset

T1	123
T2	123
T3	123
T4	12
T5	1
T6	13

TABLE II. Sensitive rule

R1	1,3 → 2

TABLE III. Association rules extracted from original dataset with MCT=0.70 and MST=0.33

Rule	Confidence	Support
2 → 1	1	0.66
2 → 3	0.75	0.50
3 → 1	1	0.66
3 → 2	0.75	0.50
2,1 → 3	0.75	0.50
3 → 1,2	0.75	0.50
1,2 → 3	0.75	0.50
1 → 3	0.66	0.66
1,3 → 2	0.75	0.50
2,3 → 1	1	0.50

TABLE IV. Genetic Algorithm parameters specifications

Population Size	20
Mutation Rate	0.01
Crossover Probability	0.80
Chromosome Length	18

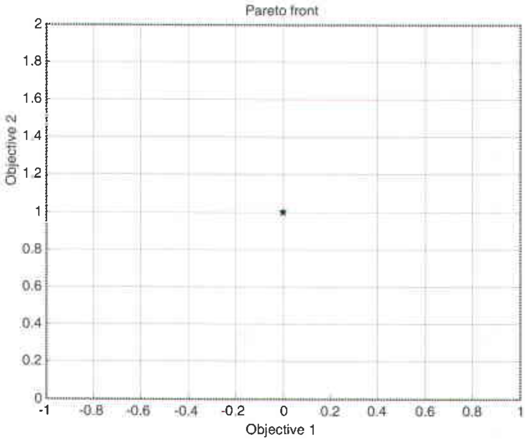

Figure 5. Pareto Front for first Fitness Function (MCT=0.70)

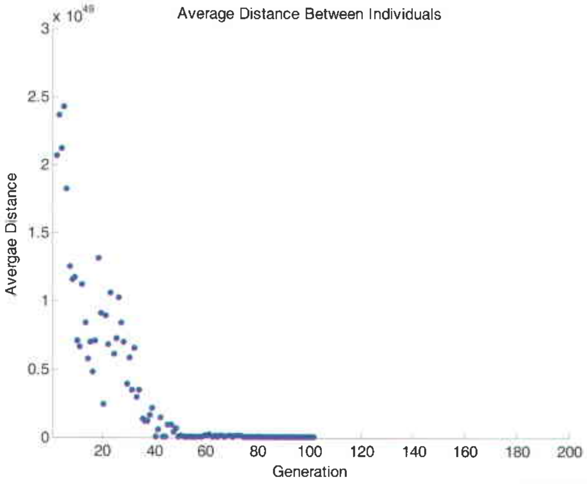

Figure 6. Average Pareto Spread for first Fitness Function (MCT=0.70)

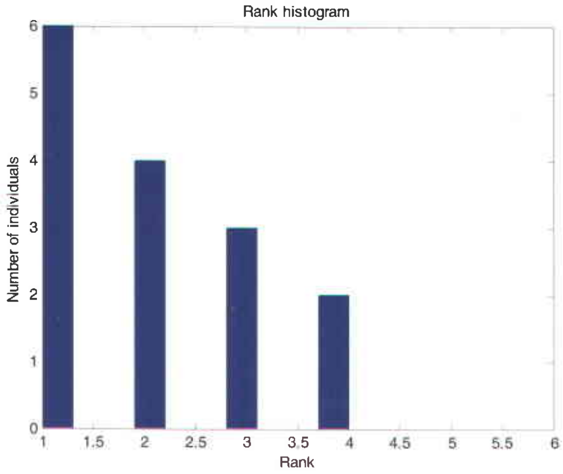

Figure 7. Ranking per number of individuals (MCT=0.70)

As we can see in figure 5, after running our method for first fitness function with Pareto ranking strategy, there is only one superior solution suggested for MCT=0.70. It means that this a best point that satisfy both objectives. In this case we should modify just one itemset to conceal the sensitive association rule. We can see the average Pareto spread for first fitness function for MCT=0.70 in figure 6. In figure 6 we can see that the average distance between individuals that the average is zero from generation 50 to 100. Ranking of individuals is depicted in figure 7.

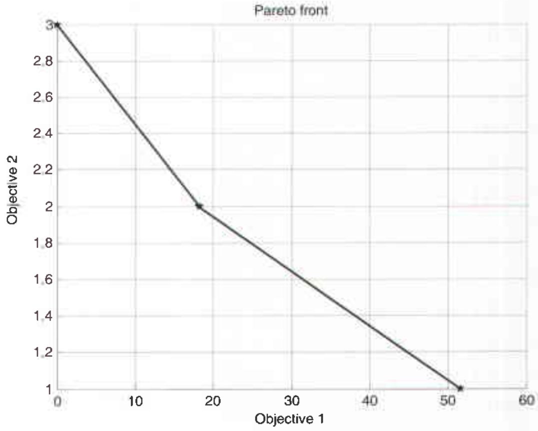

Figure 8. Pareto Front for first Fitness Function (MCT=0.15)

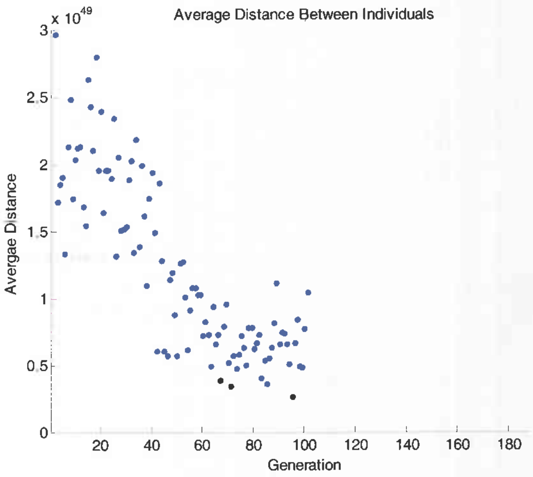

Figure 9. Average Pareto Spread for first Fitness Function (MCT=0.15)

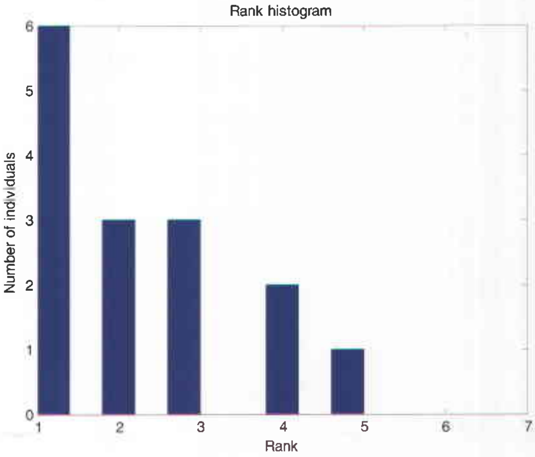

Figure 10. Ranking per number of individuals (MCT=0.15)

Figure 8 shows that by MCT=0.15 for confidence-based fitness function and with Pareto ranking strategy, there are just three superior solutions suggested for concealing sensitive rule. It means that these three points are the best points that satisfy both objectives i.e. number of modifications and hiding distance. In this case based on our priorities, we can decide that which of these sanitizations approaches should be selected and which of them should be considered more or less. If the major goal is to conceal all sensitive rules, we should select the points with less value of objective-1 than the others. If the major goal is to keep accuracy of original dataset, we should select the points with less value of objective-2. We can see the average Pareto spread for first fitness function for MCT=0.15 in figure 8. In this strategy tried to conceal sensitive rules in second phase on association rule mining (see figure 4). The average distances between individuals and ranking histogram have shown in figures 9 and 10 respectively. The variation of individual raking is higher than our prior examination.

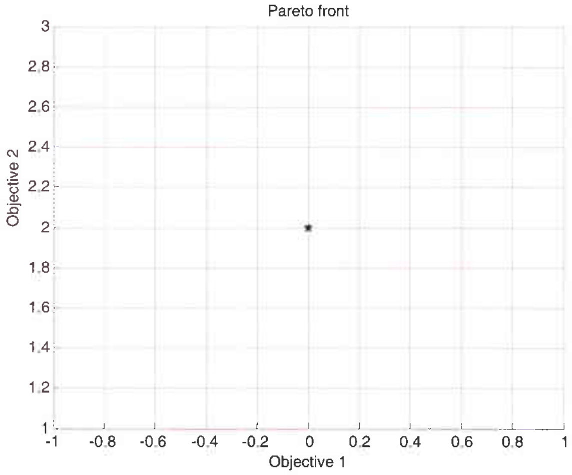

Figure 11. Pareto Front for second Fitness Function (MST=0.33)

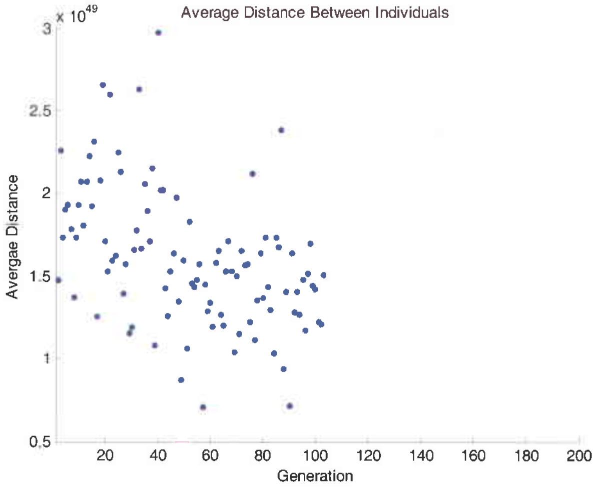

Figure 12. Average Pareto Spread for second Fitness Function (MST=0.33)

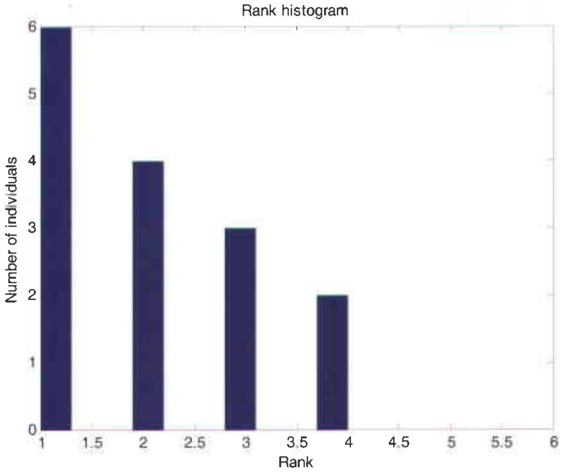

Figure 13. Ranking per number of individuals (MCT=0.33)

In figure 11 we can see that by applying support-based strategy as fitness function with Pareto ranking strategy, there is only one superior solution suggested for MST=0.33. It means that in this point both objectives are satisfied and this point dominates all other solutions. According to the solution we need modify just two itemsets to conceal the sensitive association rule. We can see the average Pareto spread values by generations for MST=0.30 in figure 12. In comparison with confidence-based fitness function, we can say that more modifications needs in this strategy because this strategy is based on support measure. Although we have more modifications in second fitness function, in this strategy tried to hide sensitive rule at higher level of security than before. There are four ranks for individual rankings in this experiment. Ranking histogram is depicted in figure 13.

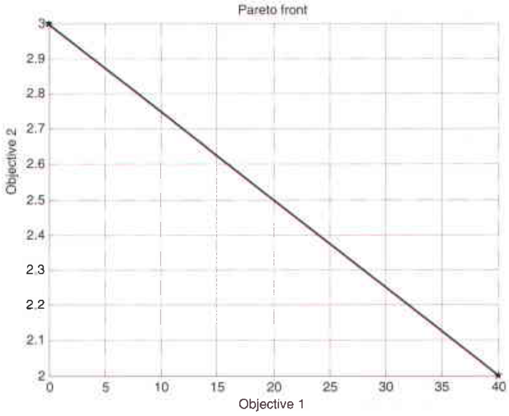

Figure 14. Pareto Front for second Fitness Function (MST=0.10)

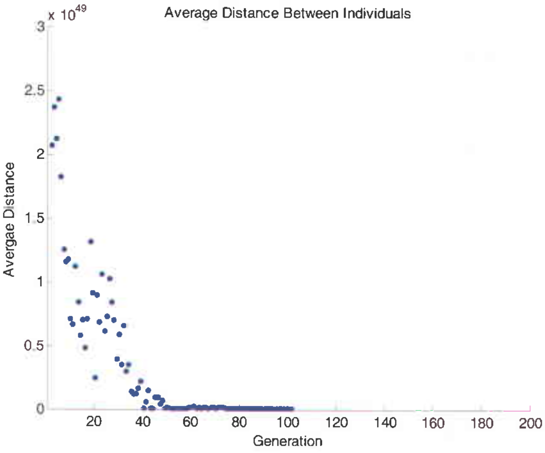

Figure 15. Average Pareto Spread for second Fitness Function (MST=0.10)

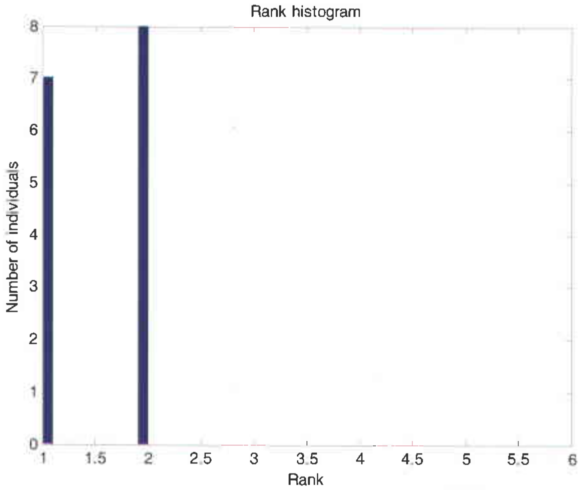

Figure 16. Ranking per number of individuals (MCT=0.10)

As depicted in figure 14, there are two superior solutions suggested for MST=0.10 which means that these points are dominating all other solutions. If the major goal is concealing all sensitive rules, we should select the points with less value of first objective and if the main goal is keeping accuracy of original dataset, we should select the points with less value of second objective. We can see the average Pareto spread diagram by first fitness function for MST=0.10 in figure 15. Remember that in this strategy tried to conceal sensitive rules in second phase on association rule mining (see figure 4). In this experiment we have only two ranks (see figure 16).

B. Computational Experiments and Results on Large Datasets

Extensive computational testing was conducted, both on real and generated datasets. This section describes the data used for computational testing, discusses the parameters used, and analyzes the results.

We have chosen *chess* and *mushroom* datasets as real-world dataset and *unknown* dataset as synthetic dataset. Characteristics of these datasets are presented in table 5.

TABLE V. Characteristics of experimental datasets

Dataset name	Number of transactions	Number of items
chess	3196	75
mushroom	8124	119
unknown	19714	194

We will present the comparison between our approach and Algorithm 1.a [3] by results obtained both on real and synthetic datasets. In our three experiments minimum confidence threshold is 5%, minimum support threshold is 7% and number of sensitive rules is chosen randomly between 5 and 10.

TABLE VI. RESULTS FOR CONFIDENCE-BASED FITNESS FUNCTION

Experiment No.	Dataset name	Number of transactions	Execution Time in sec	Number of Modifications
1	chess	3196	48.57984	5914
2	mushroom	8124	71.53492	1024
3	unknown	19714	663.4662	4

TABLE VII. RESULTS FOR ALGORITHM 1A APPROACH

Experiment No.	Dataset name	Number of transactions	Execution Time in sec	Number of Modifications
1	chess	3196	37.8613	5912
2	mushroom	8124	50.52941	1031
3	unknown	19714	510.6002	5

The results of three experiments are shown in Table 6 (for our approach) and Table 7 (for algorithm 1a). We can see that almost less number of modifications needed in our approach. As a result, the utility and accuracy of sanitized dataset keeps higher than Algo1a. On the other hand, Alog1a often has better executions time than our algorithm, because of its computational simplicity. Our approach has better performance in execution time than Algo1a, when it used for more heavy datasets. The main reason for this matter is our preprocessing phase and its good performance in preparing minimal dataset to association rule hiding. The main factor for better execution time in light datasets is that Algo1a is designed based on greedy algorithm but our approach has meta-heuristic algorithm which greedy algorithms in small solution space has better performance that other exact algorithms. Although in these three experiments greedy algorithms often have less execution time but their final solution can be non-optimal in contrast with meta-heuristic algorithm. The problem of "number of modifications" is an important issue in privacy preserving approaches. In our approach all sensitive association rule are concealed completely with modifying less number of transactions in comparison with Algo1a approach. So the accuracy of our approach is higher and we loss less number of non-sensitive rules than the other method. We can see that results in tables 6-7 support this fact.

To have overall conclusion we integrate our experiments in each dataset for different aspects. There are three key evolution factors in our sanitization research: Number of modifications, dissimilarity between original and modified dataset and execution time of sanitization approach. We present the results of experiments for chess datasets in figures 17-19. According the number of sensitive rules in sanitization process, these experiments are done there times for each method. Using this approach, we have managed to optimally solve problem that are many magnitude larger than those previously presented in the literature in terms of number of modifications, dissimilarity and execution time.

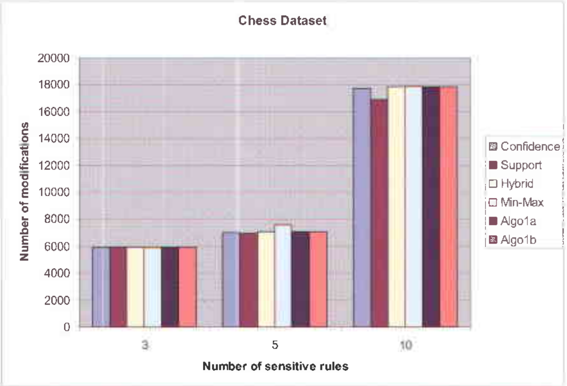

Figure 17. Number of modifications in Pareto Ranking Strategy (in four fitness function) vs. base algorithms in chess dataset

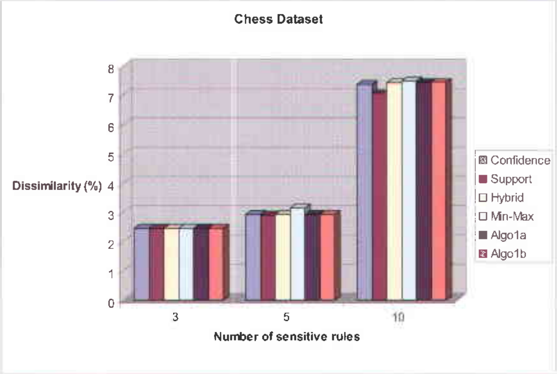

Figure 18. Dissimilarity between Pareto Ranking Strategy (in four fitness function) and base algorithms in chess dataset

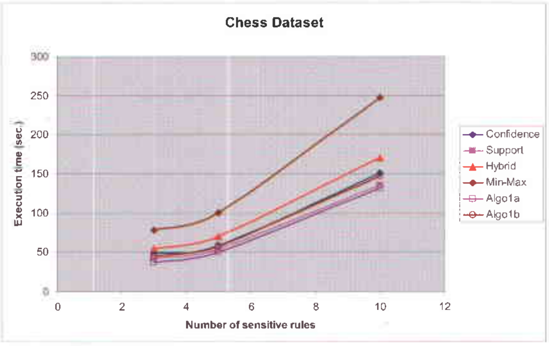

Figure 19. Execution time for Pareto Ranking Strategy (in four fitness function) and base algorithms in chess dataset

VI. Conclusions

This paper addresses the problem of hiding sensitive associations rules in transactional datasets, which is an important problem that arises when database are shared between firms. In this paper, a new multi-objective optimization algorithm is applied for privacy preserving of association rule mining. To cope with the multi-objective functions, Pareto-front ranking strategy has been applied for obtaining the non-dominated solutions front. This new method not only provides the solution efficiently, but also exposes better diversity along the Pareto-optimal front. Hence more solution choices become available for designers. Actually in this work, end-user (individual or security administrator of organization) is free to choose more interesting solutions based on her/his multi-objective priorities. This is particularly useful when proper fitness function selected for hiding and appropriate preprocessing strategy is used for concealing frequent item sets or association rules. Because of its rapid convergence capability, the proposed fitness functions have the advantage of shortening the computational time to gain the necessary results, especially by applying proper preprocessing approach in large datasets.

The key contributions in this paper can be summarized as follows: first, two pre-sanitization processes are designed. These methods select which transaction(s) and which item(s) in each transaction should be changed in order to all frequent item sets/association rules concealed safely and minimum side effect accrues. Second, four sanitization strategies proposed that comprise the hearth of our approach. Different criteria were also introduced in these sanitization strategies. The novelties of our approach are summarized in applying meta-heuristic approach for finding best solution(s), and suggesting a variety of best solutions for all objectives. Finally the work presented here introduces the idea of rule and itemset sanitization, which complements the old idea behind data sanitization. At present, we are looking for new aspects of sanitization and proposing new fitness functions according to new types of sanitization. Our permanent goal in this area is keeping privacy and accuracy of dataset as more as possible.

References

[1] Shariq J. Rizvi and Jayant R. Haritsa. Maintaining Data Privacy in Association Rule Mining. *In proceedings 28th VLDB Conference*, Hong Kong, China, 2002.

[2] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):434–447, 2004

[3] C. Clifton and D. Marks. Security and privacy implications of data mining. *SIGMOD '96: Proceedings of the 2000 ACM IGMOD International Conference on Management of Data*, pages 15–20, 1996.

[4] S. Oliveira and O. Zaiane. Privacy preserving frequent itemset mining. *RPITS'14: Proceedings of the IEEE International Conference on Privacy, Security, and DataMining*, pages 43–54, 2002.

[5] X. Sun and P. S. Yu. A border-based approach for hiding sensitive frequent itemsets. ICDM '05: Proceedings of the 5th IEEE International Conference on Data Mining, pages 426-433, 2005.

[6] M. Atallah, E. Bertino, A. Elmagarmid,M. Ibrahim and V. Verykios. Disclosure limitation of sensitive rules. *Proc. of IEEE Knowledge and Data Engineering Exchange Workshop (KDEX)*, November 1999

[7] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni. Association Rule Hiding. *IEEE Trans. on Knowledge and Data Engineering*, 16(4), 2004.

[8] S. Oliveira and O. Zaiane. Privacy preserving frequent itemset mining. *CRPITS'14: Proceedings of the IEEE International Conference on Privacy, Security, and Data Mining*, pages 43–54, 2002.

[9] L. David, *Handbook of Genetic Algorithms*. New York : Van Nostrand Reinhold. 1991.

[10] D.E. Goldberg, *Genetic Algorithms: in Search, Optimization, and Machine Learning*. New York : Addison-Wesley Publishing Co. Inc. 1989.

[11] D. Goldberg, B. Karp, Y. Ke, S. Nath, and S. Seshan, *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989.

[12] I.Y. Kim and O.L. de Weck, Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Struct Multidisc Optim. 29*, 149–158, *Springer*, 2005.

[13] A. Amiri. Dare to share: Protecting sensitive knowledge with data sanitization. *Decision Support Systems*, 43(1):181–191, 2007.

[14] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005)*, pages 466–473, 2005.

[15] S.-L.Wang and A. Jafari. Using unknowns for hiding sensitive predictive association rules. In *Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration (IRI 2005)*, pages 223–228, 2005.

[16] X. Wu, Y. Wu, Y. Wang, and Y. Li. Privacy aware market basket data set generation: A feasible approach for inverse frequent set mining. In *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM 2005)*, 2005.

[17] Y.-H.Wu, C.-M. Chiang, and A. L. P. Chen. Hiding sensitive association rules with limited side effects. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):29–42, 2007.

[18] O. Abul, M. Atzori, F. Bonchi, and F. Giannotti. Hiding sequences. Technical report, Pisa KDD Laboratory, ISTI-CNR, Area della Ricerca di Pisa, Nov. 2006.

[19] A. Gkoulalas-Divanis and V. S. Verykios. An integer programming approach for frequent itemset hiding. In *Proceedings of the 2006 ACM Conference on Information and Knowledge Management (CIKM 2006)*, pages 748–757, 2006.

[20] A. `Inan and Y. Saygin. Privacy preserving spatio-temporal clustering on horizontally partitioned data. In *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2006)*, pages 459–468, 2006.

[21] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright. A new privacy preserving distributed k-clustering algorithm. In *Proceedings of the 2006 SIAM International Conference on Data Mining (SDM 2006)*, 2006.

Mohammad Naderi Dehkordi was born in Isfahan, Iran, in 1977. He has a Bachelor's degree in Computer Engineering from Isfahan University of Technology, Isfahan, Iran and a Master's degree in Computer Engineering from Najafabad Branch, Islamic Azad University. He has received his Ph.D. in Computer Engineering from Science and Research Branch, Islamic Azad University, Tehran, Iran, majoring in Privacy Preserving Data Mining. His main research interests include On-Line Analytical Processing and engineering privacy preserving in Hippocratic databases.

Kambiz Badie has received his B.Sc., M.Sc. and Ph.D. in Electronic Engineering from the Tokyo Institute of Technology, Japan, majoring in Pattern Recognition & Artificial Intelligence. His major research interests are Machine Learning, Cognitive Modeling, and Systematic Knowledge Processing in general, and Analogical Knowledge Processing, Experience-Based Modeling, and Interpretative Modeling in particular with emphasis on Idea , Technique, and Content Generation. He has published more than 150 conference & 30 journal papers in this regard. His recent published books are "Clustering" and also "Strangification: A New Paradigm in Knowledge Processing & Creation".

He has been involved in many scientific and managerial positions such as Head of IT Research Faculty at Iran Telecom Research Center during these years. At present, he is a member of scientific board of Iran Telecom Research Center and University of Tehran.

Ahmad Khadem-Zadeh was born in Meshed, Iran, in 1943. He received the B.Sc. degree in applied physics from Ferdowsi University, Meshed, Iran, in 1969 and the M.Sc., Ph.D. degrees respectively in Digital Communication and Information Theory & Error Control Coding from the University of Kent, Canterbury, UK.

He is currently the Head of Education & National Scientific and International Scientific Cooperation Department at Iran Telecom Research Center (ITRC). He was the head of Test Engineering Group and the director of Computer and Communication Department at ITRC. He is also a lecturer at Tehran Universities & he is a committee member of the Iranian Electrical Engineering Conference Permanent Committee. Dr. Khadem-Zadeh has been received four distinguished national and international awards including Kharazmi International Award, and has been selected as the National outstanding researcher of the Iran Ministry of Information and Communication Technology.